SYSTRAN's Chinese Word Segmentation

Jin Yang
SYSTRAN Software, Inc.
9333 Genesee Ave.
San Diego, CA 92121, USA
jyang@systransoft.com

Jean Senellart
SYSTRAN S.A.
1, rue du Cimetière
95230 Soisy-sous-Montmorency, France
senellart@systran.fr

Remi Zajac
SYSTRAN Software, Inc.
9333 Genesee Ave.
San Diego, CA 92121, USA
zajac@systransoft.com

Abstract

SYSTRAN's Chinese word segmentation is one important component of its Chinese-English machine translation system. The Chinese word segmentation module uses a rule-based approach, based on a large dictionary and fine-grained linguistic rules. It works on general-purpose texts from different Chinese-speaking regions, with comparable performance. SYSTRAN participated in the four open tracks in the First International Chinese Word Segmentation Bakeoff. This paper gives a general description of the segmentation module, as well as the results and analysis of its performance in the Bakeoff.

1 Introduction

Chinese word segmentation is one of the pre-processing steps of the SYSTRAN Chinese-English Machine Translation (MT) system. The development of the Chinese-English MT system began in August 1994, and this is where the Chinese word segmentation issue was first addressed. The algorithm of the early version of the segmentation module was borrowed from SYSTRAN's Japanese segmentation module. The program ran on a large word list, which contained 600,000 entries at the time[1]. The basic strategy was to list all possible matches for an entire linguistic unit, then solve the overlapping matches via linguistic rules. The development was focused on technical domains, and high accuracy was achieved after only three months of development. Since then, development has shifted to other areas of Chinese-English MT, including the enrichment of the bi-lingual word lists with part-of-speech, syntactic and semantic features. In 2001, the development of a prototype Chinese-Japanese MT system began. Although the project only lasted for three months, some important changes were made in the segmentation convention, regarding the distinction between words and phrases[2]. Along with new developments of the SYSTRAN MT engine, the segmentation engine has recently been re-implemented. The dictionary and the general approach remain unchanged, but dictionary lookup and rule matching were re-implemented using finite-state technology, and linguistic rules for the segmentation module are now expressed using a context-free-based formalism, improving maintainability. The re-implementation generates multiple segmentation results with associated probabilities. This will allow for disambiguation at a later stage of the MT process, and will widen the possibility of word segmentation for other applications.

2 System Description

2.1 Segmentation Standard

Our definition of words and our segmentation conventions are based on available standards, modified for MT purposes. The PRC standard (Liu et al., 1993) was initially used. Sample differences are listed as follows:

Type	PRC	SYSTRAN
NP	中华民族 中华人民共和国	中华 民族 中华 人民 共和国
CD	31 日	31 日
CD + M	一个 一排排	一 个 一 排 排
DI4 + CD	第一	第 一
Name	李 白 李 清照	李白 李清照

Table 1. Segmentation Divergences with the PRC Guidelines

2.2 Methodology

The SYSTRAN Chinese word segmentation module uses a rule-based approach and a large dictionary. The dictionary is derived from the

Chinese-English MT dictionary. It currently includes about 400,000 words. The basic segmentation strategy is to list all possible matches for a translation unit (typically, a sentence), then to solve overlapping matches via linguistic rules. The same segmentation module and the same dictionary are used to segment different types of text with comparable performance.

All dictionary lookup and rule matching are performed using a low level Finite State Automaton library. The segmentation speed is 3,500 characters per second using a Pentium 4 2.4GHZ processor.

Dictionary

The Chinese-English MT dictionary currently contains 400,000 words (e.g., 中华), and 200,000 multi-word expressions (e.g., 中华 人民 共和国). Only words are used for the segmentation. Specialized linguistic rules are associated with the dictionary. The dictionary is general purpose, with good coverage on several domains. Domain-specific dictionaries are also available, but were not used in the Bakeoff.

The dictionary contains words from different Chinese-speaking regions, but the representation is mostly in simplified Chinese. The traditional characters are considered as "variants", and they are not physically stored in the dictionary. For example, 意大利 and 义大利 are stored in the dictionary, and 義大利 can also be found via the character matching 義→ 义.

The dictionary is encoded in Unicode (UTF8), and all internal operations manipulate UTF8 strings. Major encoding conversions are supported, including GB2312-80, GB13000, BIG-5, BIG5-HKSCS, etc.

Training

The segmentation module has been tested and fine-tuned on general texts, and on texts in the technical and military domains (because of specific customer requirements for the MT system). Due to the wide availability of news texts, the news domain has also recently been used for training and testing.

The training process is merely reduced to the customization of a SYSTRAN MT system. In the current version of the MT system, customization is achieved by building a User Dictionary (UD). A UD supplements the main dictionary: any word that is not found in the main MT system dictionary is added in a User Dictionary.

Name-Entity Recognition and Unknown Words

Name entity recognition is still under development. Recognition of Chinese persons' names is done via linguistic rules. Foreign name recognition is not yet implemented due to the difficulty of obtaining translations.

Due to the unavailability of translations, even when an unknown word has been successfully recognized, we consider the unknown word recognition as part of the terminology extraction process. This feature was not integrated for the Bakeoff.

2.3 Evaluation

Our internal evaluation has been focused on the accuracy of segmentation using our own segmentation standard. Our evaluation process includes large-scale bilingual regression testing for the Chinese-English system, as well as regression testing of the segmenter itself using a test database of over 5MB of test items. Two criteria are used:

1. Overlapping Ambiguity Strings (OAS): the reference segmentation and the segmenter segmentation overlap for some string, e.g., AB-C and A-BC. As shown below, this typically indicates an error from our segmenter.

2. Covering Ambiguity Strings (CAS): the test strings that cover the reference strings (CAS-T: ABC and AB-C), and the reference strings that cover the test strings (CAS-R: AB-C and ABC). These cases arise mostly from a difference between equally valid segmentation standards.

No evaluation with other standards had been done before the Bakeoff.

Test	Reference	Type
崇文区　政府	崇文　区政府	OAS
冰清玉洁	冰 清 玉 洁	CAS-T
除夕之夜	除夕 之 夜	CAS-T
擦泪	擦 泪	CAS-T
精神 文明	精神文明	CAS-R
1994　年	1994 年	CAS-R
不 怕	不怕	CAS-R

Table 2. Types of Segmentation Differences

3 Discussion of the Bakeoff

3.1 Results

SYSTRAN participated in the four open tracks in the First International Chinese Word Segmentation Bakeoff http://www.sighan.org/bakeoff2003/. Each track corresponds to one corpus with its own word segementation standard. Each corpus had its own segmentation standard that was significantly different from the others. The training process included building a User Dictionary that contains words found in the training corpora, but not in the SYSTRAN dictionary. Although each of these corpora was segmented according to its own standard, we made a single UD containing all the words gathered in all corpora.

Although the ranking of the SYSTRAN segmenter is different in the four open tracks, SYSTRAN's segmentation performance is quite comparable across the four corpora. This is to be compared to the scores obtained by other participants, where good performance was typically obtained on one corpus only. SYSTRAN scores for the 4 tracks are shown in Table 3 (Sproat and Emerson, 2003).

Track	R	P	F	R_{oov}	R_{iv}
AS_o	0.915	0.894	0.904	0.426	0.926
CTB_o	0.891	0.877	0.884	0.733	0.925
HK_o	0.898	0.860	0.879	0.616	0.920
PK_o	0.905	0.869	0.886	0.503	0.934

Table 3. SYSTRAN's Scores in the Bakeoff

3.2 Discussions

The segmentation differences between the reference corpora and SYSTRAN's results are further analyzed. Table 4 shows the partition of divergences between OAS, CAS-T, and CAS-R strings:[3]

	Total	Same	OAS	CAS-T	CAS-R
AS_o	11,985	10,970	76	448	491
CTB_o	39,922	35,561	231	2,419	1,711
HK_o	34,959	31,397	217	1,436	1,909
PK_o	17,194	15,554	82	615	943

Table 4. Count of OAS and CAS Divergence

The majority of OAS divergences show incorrect segmentation from SYSTRAN. However, differences in CAS do not necessarily indicate incorrect segmentation results. The reasons can be categorized as follows: a) different segmentation standards, b) unknown word problem, c) name entity recognition problem, and d) miscellaneous[4]. The distributions of the differences are further analyzed in Table 5 and 6 for the AS_o and PK_o corpora, respectively.

CAS-R: Unique Strings=334 (total=491)			
Type	Count	Percent	Examples
Different Standards	184	55%	感觉到 不能 第十三区 廿十五日
Unknown Words	116	35%	秋颱 中菜 哭骂 院庆
Name Entity	30	9%	川崎 津巴貝 台塑
Misc.	4	1%	一百余萬
CAS-T: Unique Strings=137 (total=448)			
Type	Count	Percent	Examples
Different Standard	134	98%	喝酒 出了名 喝不喝酒
True Covering	3	2%	都會 有為

Table 5. Distribution of Divergences in the AS_o Track

CAS-R: Unique Strings=508 (total=943)			
Type	Count	Percent	Examples
Different Standards	294	58%	中共中央 这次 本届 不要 第一 2001 年
Unknown Words	90	18%	攀岩 雪浴 拥堵
Name Entity	61	12%	奥佩蒂 福彩村
Misc.	63	12%	20% 3.9 亿
CAS-T: Unique Strings=197 (total=615)			
Type	Count	Percent	Examples
Different Standards	194	98%	中国人 大吼 不夜天 赤着膊
True Covering	3	2%	高过 雪洗

Table 6. Distribution of Divergences in the PK_o Track

This analysis shows that the segmentation results are greatly impacted by the difference in the segmentation standards. Other problems include for example the encoding of numbers using single bytes instead of the standard double-byte encoding in the PKo corpus, which account for about 12% of differences in the PKo track scores.

4 Conclusion

For an open track segmentation competition like the Bakeoff, we need to achieve a balance between the following aspects:

- Segmentation standards: differences between one's own standard and the reference standard.
- Adaptation to the other standards: whether one should adapt to other standards.
- Dictionary coverage: the coverage of one's own dictionary and the dictionary obtained by training.
- Algorithm: combination of segmentation, unknown word identification, and name entity recognition.
- Speed: the time needed to segment the corpora.
- Training: time and manpower used for training each corpus and track

Few systems participated in all open tracks: only SYSTRAN and one university participated in all four. We devoted about 2 person/week for this evaluation. We rank in the top three of three open tracks, and only the PK$_o$ track scores are lower, probably because of encoding problems for numbers for this corpus (we did not adjust our segmenter to cope with this corpus-specific problem). Our results are very consistent for all open tracks, indicating a very robust approach to Chinese segmentation.

Analysis of results shows that SYSTRAN's Chinese word segmentation excels in the area of dictionary coverage, robustness, and speed. The vast majority of divergences with the test corpora originate from differences in segmentation standards (over 55% for CAS-R and about 98% for CAS-T). True errors range between 0% and 2% only, the rest being assigned to either the lack of unknown word processing or the lack of a name entity recognizer. Although not integrated, the unknown word identification and name entity recognition are under development as part of a terminology extraction tool.

For future Chinese word segmentation evaluations, some of the issues that arose in this Bakeoff would need to be addressed to obtain even more significant results, including word segmentation standards and encoding problems for example. We would also welcome the introduction of a surprise track, similar to the surprise track of the DARPA MT evaluations that would require participants to submit results within 24 hours on an unknown corpus.

References

Liu, Y, Tan Q. & Shen, X. 1993. Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology.

Sproat, R., & Emerson T. 2003. The First International Chinese Word Segmentation Bakeoff. In the Proceedings of the Second SIGHAN Workshop on Chinese Language Processing. ACL03.

[1] The word list only contained Chinese-English bilingual dictionary without any syntactic or semantic features. It also contained many compound nouns, e.g. 北京大学.

[2] Compound nouns are no longer considered as words. They were moved to the expression dictionary. For example, 北京大学 has become 北京 大学.

[3] The number of words in the reference strings is used when counting OAS and CAS divergences. For example, 除夕之夜's CAS count is three because the number of words in the reference string 除夕 之 夜 is three.

[4] Word segmentation in SYSTRAN MT systems occurs after sentence identification and normalization. During word segmentation, Chinese numbers are converted into Arabic numbers.